事务型数据库系统中的数据拷贝过去。这个向数据仓库拷贝数据的过程被称为提取 - 转换 - 加载（Extract-Transform-Load，ETL）。ETL 的基本流程是：从事务型数据库中提取数据，将其转换为通用表示形式（可能包含数据验证、数据归一化、编码、去重、表模式转换等工作），最终加载到分析型数据库中。该流程可能会非常麻烦，通常需要复杂的技术方案来满足性能要求。为了保持数据仓库中的数据同步，ETL 过程需要周期性地执行。

一旦数据导入数据仓库，我们就能对它们做查询分析。通常数据仓库中的查询可以分为两类：第一类是定期报告查询。它可用于计算业务相关的统计数据，如收入、用户增长、产出等。将这些指标整合成报告，能够帮助管理层评估企业整体健康状况。第二类是即席查询（ad-hoc query）。其主要目的是通过解答特定问题来辅助关键性的商业决策，例如通过查询来整合营收数字和电台广告中的投入，以评估市场营销的有效性。如图 1-3 所示，无论哪一类查询，都是在数据仓库中以批处理的方式执行。

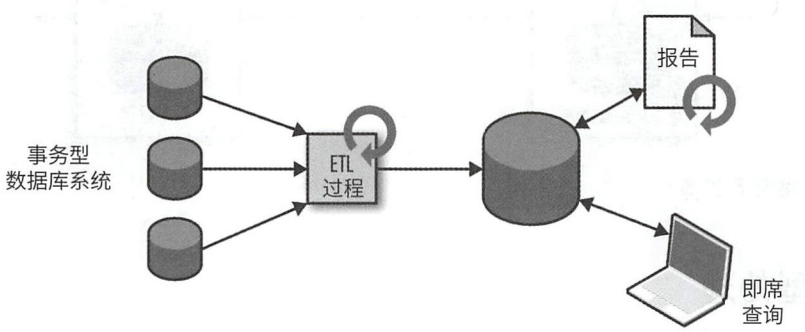

图 1-3：用于数据分析的传统数据仓库架构

时至今日，Apache Hadoop 生态组件已经成为很多公司和企业 IT 基础设施中举足轻重的部分。海量日志文件、社交媒体、网页点击日志等数据已不再使用关系数据库系统存储，而是会写入 Hadoop 分布式文件系统（HDFS）、S3 或其他诸如 Apache HBase 的批量数据存储系统。这些系统以低廉的成本提供庞大的存储容量，而它们中的数据也可以通过很多基于 Hadoop 的 SQL 引擎（如